Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 202–216
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Non-linear correlation of content and metadata information extracted from biomedical article datasets

Theodosios Theodosiou *, Lefteris Angelis, Athena Vakali

Department of Informatics, School of Natural Sciences, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Received 30 December 2006
Available online 10 June 2007

Abstract

Biomedical literature databases constitute valuable repositories of up to date scientific knowledge. The development of efficient machine learning methods in order to facilitate the organization of these databases and the extraction of novel biomedical knowledge is becoming increasingly important. Several of these methods require the representation of the documents as vectors of variables forming large multivariate datasets. Since the amount of information contained in different datasets is voluminous, an open issue is to combine information gained from various sources to a concise new dataset, which will efficiently represent the corpus of documents. This paper investigates the use of the multivariate statistical approach, called Non-Linear Canonical Correlation Analysis (NLCCA), for exploiting the correlation among the variables of different document representations and describing the documents with only one new dataset. Experiments with document datasets represented by text words, Medical Subject Headings (MeSH) and Gene Ontology (GO) terms showed the effectiveness of NLCCA.

Keywords: Biomedical literature; Non-linear canonical correlation; MeSH; Gene Ontology; Data reduction; Visualization; PubMed

1. Introduction

Biomedical literature is stored in large-scale organized repositories which tend to exhibit increasing demands in terms of space and access speeds since current scientific biological knowledge is emerging and evolving. The manual management of this information and the extraction of useful scientific knowledge is a tedious and laborious work that involves the identification of pertinent literature and data from a large number of articles [1]. For example, the PubMed database contains over 14 million bibliographic citations and abstract articles from more than 4800 journals [2]. In this context, clustering and classification methods are becoming increasingly necessary for efficiently organizing, exploring and retrieving the information of biomedical literature.

There are several research efforts in the machine learning field focusing on the clustering or classification of biomedical documents from PubMed and on the extraction of useful biological knowledge [7–12]. Some of the proposed methods are maximum-entropy [8], linear discriminant analysis [9], Naïve Bayes [10], support vector machines [11], group-average agglomerative clustering [12], etc.

The main characteristic of the methods used so far is the numeric representation of biomedical documents. By numeric representation we refer to having each document represented as a vector of numeric variables with each variable corresponding to a (informative/important) word in the document. Similar representations have appeared using vectors with variables (metadata) which correspond to specific terms from a structured controlled vocabulary usually created manually for concisely describing specific information and knowledge [3–5].

Two popular metadata approaches are the Medical Subject Headings (MeSH) [5] and the Gene Ontology (GO)

* Corresponding author. Fax: +30 2310 998419.
E-mail address: theodos@csd.auth.gr (T. Theodosiou).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.06.004

[3,4] which are used for information retrieval from biomedical document databases:

- MeSH is mainly used by the PubMed database for semantically relating documents in order to enhance the search and improve the accuracy of the query system of PubMed[1] [6]. The MeSH terms are arranged both alphabetically and hierarchically. The hierarchy has 11 levels. The upper levels of the hierarchical structure contain the most general MeSH terms, for example ''Diseases'' appear in level one, whereas more specific terms, like ''Mitochondrial myopathies'' are found in lower levels. Each PubMed document usually has 10–12 MeSH terms, manually assigned to it by experts that read the full-length document [5].
- GO is mainly used for describing gene products (RNA or proteins) in a species independent manner, but it has also been used for describing the information contained in biomedical documents. The GO terms are structured as nodes in a Direct Acyclic Graph (DAG). They are divided into three subgroups (sub-ontologies): biological processes, cellular components and molecular functions [3,4]. The GO is not part of the PubMed system, but it has been used in several research work for describing biomedical documents (e.g. [1,7–12]).

It should be noted that some terms are found in both MeSH and GO, for example ''DNA Methylation'' or ''Autophagy'' appear in both MeSH and GO terms listings.

A common characteristic of most of the research efforts is that the data representation of the documents is based on the Vector Space Model (VSM) [13–15] which transforms a document into a vector of numeric variables. The variables may represent words of the documents [8,9], MeSH terms [10] or GO terms [8,9,11]. Each of the aforementioned representations describes differently the scientific information contained in the documents. For example, the MeSH terms can be more general than the GO terms about the molecular concepts explained in the document, since GO terms are specifically built for describing molecular concepts. This fact raises a question of how the different representations are correlated and whether they can be combined in a smaller and more concise dataset.

Usually the number of variables is quite large and thus various efforts [9,10,12] deal with the dimensionality reduction of the dataset describing the documents. Certain statistical methods, like Singular Value Decomposition [12] or Principal Component Analysis [9], exploit the correlation structure of the dataset in order to produce new variables, fewer and uncorrelated [12]. This is advantageous for other subsequent procedures, like certain classification methods, which assume uncorrelated variables. Nevertheless, the obvious fact is that the words of a document's text, the MeSH or the GO terms are semantically related. For example, the possibility of finding in the same document the words ''Methylation'' and ''Cancer'' is higher, than finding ''Methylation'' and ''Chemimechanical coupling'' together.

Since data reduction is important in order to reveal patterns, trends or groupings in biomedical documents, in [9] the authors have considered a classification (discriminant analysis) method which produces new and fewer variables, able to graphically represent the documents in a low-dimensional Euclidean space. This method models the relationships between words in the text and corresponding GO terms of biomedical articles. The results showed that the reduced number of new variables can efficiently describe the documents and compete with other classification methods, such as Support Vector Machines that use all the original variables.

The motivation of the present work is to consider all of the above document representations and propose a method which would combine them, by exploiting the information they contain. The methodology under consideration is the Non-Linear Canonical Correlation Analysis (NLCCA) [16,17] which can combine information from more than two document representations by exploiting their correlation structure. The method's general principle is the optimal projection of the documents on a low-dimensional Euclidean space by computing a set of scores from the original variables in such a way that the new scores can adequately explain the variance of the original data, i.e. document words, MeSH and GO terms. It must be noted that NLCCA is not restricted to linear relationships and is especially useful for analyzing all kinds of data, numerical and categorical [22]. In general, the contribution of NLCCA to the text mining problems addressed so far by earlier efforts can be considered to be the abridgement of information from two or more datasets describing the same documents.

The rest of the paper is organized as follows: Section 2 describes the document representations used for the experimentation. Section 3 outlines NLCCA and various measures for evaluating the method. Section 4 presents the results of the experiments. Section 5 provides practical conclusions regarding the applicability of NLCCA and discusses future work.

2. Document representation

This section provides an outline of all the techniques leading to the representation of the documents as numerical vectors. Words, MeSH and GO terms are transformed into variables, in order to apply any statistical method to them.

We consider a specific set of biomedical documents which is extracted from the PubMed database, using its query system, where each query is relevant to a specific GO term (Fig. 1). The documents retrieved contain the title, the abstract, other bibliographic citations (authors' names, etc.) and the MeSH terms assigned to them by

[1] Part of the National Library of Medicine of the National Institutes of Health in the U.S.A.

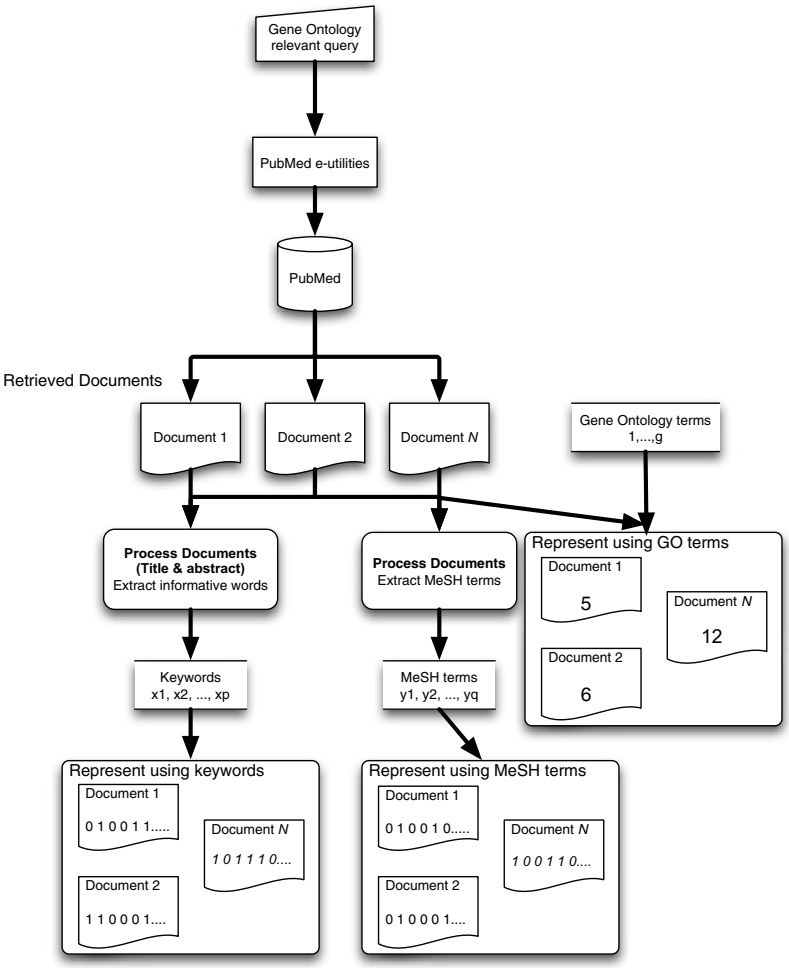

Fig. 1. The processes involved in retrieving the documents and representing them with three different matrices using the keywords, the MeSH and the GO terms.

the PubMed experts. Only the words from the title, the abstracts and the MeSH terms were used for the subsequent steps and the analysis.

The information contained in the documents is transformed to numerical vectors, following the most common (in text mining) document representation model, known as Vector Space Model (VSM) [13–15]. The most informative and important (with respect to the document's content) words are extracted from the title and the abstract of each document by a procedure involving the typical following steps [9,15,18]:

Tokenization. Extraction of all words (case insensitive) from the entire set of documents;
Stopword removal. Elimination of non-content-bearing, non informative words, called ''stopwords'' such as ''a'', ''and'', ''the'', etc. The removed stopwords are the same to the ones listed by PubMed[2];
Stemming. Using only the root of each word, e.g. exclude 'ing' from 'processing'. The algorithm we used for stemming is based on [19] and is implemented in Perl by Mary D. Taffet[3];
Frequency count. Counting the number of occurrences of each word in each document;
Filtering. Exclusion of ''high-frequency'' and ''low-frequency'' words. The filtering of the words of a document is a common practice in biological text mining, aiming to remove non-informative words [8,20]. Words with too high frequency (i.e. occurring to almost all of the documents) or too infrequent words (occurring to only a very small number of documents) are not descriptive of the document corpus and thus can be excluded from the processing. In our analysis we decided to exclude the words appearing in more than 95% of the total number of articles and also those appearing in less than 0.05% of the total number of articles. These cutoff values were empirically determined, based on experiments from

[2] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.table.pubmedhelp.T43.

[3] http://www.comp.lancs.ac.uk/computing/research/stemming/Files/Perl.zip.

our previous published work [9] and are quite similar to those used in other works such as [8] and [20]. The remaining words after filtering are called *keywords* in the rest of the paper.

In Fig. 2 we can see schematically an example of applying the above steps to a biomedical document. The document in the first frame of Fig. 2 contains a single sentence just for illustrative purposes and is relevant to the biological phenomenon of autophagy. The frequency of the words is imaginary and it is used to illustrate the filtering step. The first step involves the identification of the words (tokens). Then the stopwords are removed, like "of" and "in". After that, the root of each word is identified, e.g. "involvement" becomes "involv". Finally, the frequency of each word is counted and the ones that are too common or too rare are excluded, like "involv" whose percentage in the total number of articles is 98%.

Having the resulted keywords extracted from all the retrieved documents dataset, we model each document by a vector of weights which express the relation of the particular document with a particular keyword. The optimal weighting scheme depends on the relevant articles for each GO code [11], and it is not possible to be known a priori [21]. The simplest weighting policy is to denote by "1" the appearance of a keyword in the document and by "0" its absence. Although there are various other more complicated weighting schemes [15,18,21] they do not guarantee better performance [21]. In [21], the authors have conducted an extensive study with different weighting schemes, i.e. binary, TF.IDF, etc. and they concluded that the best weighting scheme for a given data set can not be known a priori. Simple weighting schemes like the binary one have been found to work very well compared to more complex ones. Since it is not guaranteed that complex weighting schemes can describe better our data and in order to avoid cumbersome calculations we used the binary weighting.

Therefore, for each document's representation we need to specify whether the particular keywords are included in the document or not, so we propose a binary weighting function as follows:

$$x_{di} = \begin{cases} 1 & \text{if keyword}_i \in d \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, p \tag{1}$$

where x_{di} refers to the presence of keyword i in document d, for a total number of p keywords.

Similarly, we define a binary weighting scheme for representing the MeSH terms within each of the documents:

$$y_{di} = \begin{cases} 1 & \text{if MeSH term}_i \in d \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, q \tag{2}$$

where the y_{di} refers to the presence of MESH term i in document d, for a total number of q MeSH terms.

After applying the above schemes, each document is finally represented by three different data structures. Specifically, for each document we have:

1. a p-dimensional vector $(x_{d1}, x_{d2}, \ldots, x_{dp})$, of keyword weights;
2. A q-dimensional vector $(y_{d1}, y_{d2}, \ldots, y_{dq})$ of MeSH term weights;
3. A single GO term which for simplicity is denoted by a number in $\{1, \ldots, g\}$ where g is the total number of GO terms. The reason for using such a single number is that in our analysis each document is assigned to only one GO term. However, if we decide to work with documents which are assigned to more than one GO terms, a similar weighting scheme as (1) and (2) can be used.

Finally, from the documents datasets we result in a corpus of N documents which are represented in a threefold manner, namely, an $N \times p$ matrix $\mathbf{X}$ (corresponding to keywords), an $N \times q$ matrix $\mathbf{Y}$ (corresponding to MeSH terms) and an $N \times 1$ matrix (i.e. a column vector) $\mathbf{z}$ (corresponding to GO terms). Note that the values of all those parameters in our experiments are: $N = 9009$, $p = 1642$, $q = 50$ and $g = 12$. A detailed description of our dataset is given in Section 4. Fig. 3 shows an example of the three different document representations.

3. The Non-Linear Canonical Correlation Analysis (NLCCA)

The Non-Linear Canonical Correlation Analysis (NLCCA) is a multivariate statistical technique aiming to explore and model the strength of the correlation between two or more datasets. Moreover, NLCCA is also a data reduction method in the sense that it combines the original variables in order to produce a new dataset with fewer variables, called scores. This is attained by scaling and projecting the initial datasets as points in a low-dimensional Euclidean space [16,17].

Since any data reduction technique involves loss of information, the mathematical objective is to attain a pro-

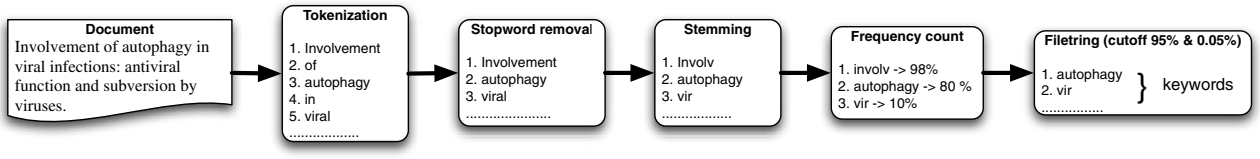

Fig. 2. An example of the steps required to extract keywords from a document.

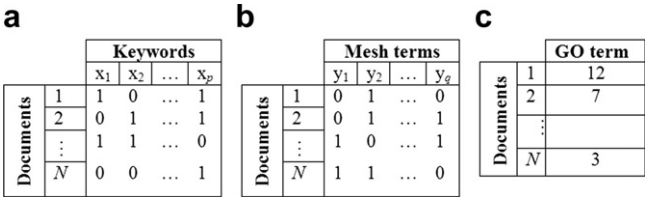

Fig. 3. Different document representations: (a) documents as a matrix (**X**) of keywords, (b) Documents as a matrix (**Y**) of MeSH terms and (c) documents as a matrix **z** (column vector) of GO terms.

jection, i.e. a set of new variables that minimizes a certain loss function. This loss function is essentially a distance between original and transformed data. Therefore, the method involves an algorithm that solves an optimization problem. The solution of the problem is a $N \times r$ matrix which in effect represents each one of the Ndocuments by a vector of r scores. The number of new dimensions r is defined by the user and for graphical purposes is usually two or three, while the maximum number of new dimensions $r_{\max}$ that can be computed by the algorithm depends on the number of datasets, the types and the number of variables in each dataset [22]. In our experiments we used the maximum number of dimensions that NLCCA can produce. From the maximum number of dimensions we can choose to work with fewer, when certain evaluation measures that are described later in the section (like the eigenvalues) indicate that some dimensions can be omitted as unimportant.

The optimization problem was first described in the Gifi system [16] and in [17]. It can be solved computationally efficiently by the Alternating Least Squares algorithm [16]. The algorithm is implemented in the SPSS statistical software, which we applied to the data of our study [22].

The projection of the documents on a Euclidean space is particularly useful for visualizing their biomedical information and for exploring patterns, groupings and for finding outliers. Intuitively, we expect that content-relevant documents will be represented by points closely positioned in the new Euclidean space, while irrelevant documents will be far away. Another appealing aspect of the method is that the new scores are uncorrelated and ranked according to their importance for explaining the overall correlation between the different document representations. So, the first dimension encloses the most significant part of the information from the original datasets, the second dimension contains the second significant part and so on. This property is particularly useful for data visualization, in the sense that we can draw important conclusions regarding the data structure by using the two or three first dimensions for plotting the documents.

There are certain statistical measures computed by NLCCA to assess the correlation between different sets and the efficiency of the new representation [22]. The fit and loss measures indicate the capability of the NLCCA solution to fit on the optimally scaled data with respect to the correlation between the datasets. The output of the procedure contains the fit value, various loss values, and the so-called eigenvalues.

Loss values are computed separately for each new dimension and each set. Every loss value represents the proportion of variation in the object scores that cannot be explained by the combination of the variables in the specific set. An average loss over all sets is also computed. A small loss value in each new dimension signifies that this dimension is capable for describing a significant amount of variation of our data. Also a small average loss value suggests that overall the new dimensions can efficiently replace the original variables of the data. In other words, the loss value indicates the loss of information we can anticipate on the new data due to the reduction of the original dimensions. For example, the average loss value for the correlation between the keywords and the MeSH terms is 11.253 (see Section 4.2), showing that the loss of information from reducing the original dimensions from 1692 to 50 is only 22.5% (computed as a ratio of 11.253/50).

The *fit value* equals the number of new dimensions minus the average loss. In the trivial case where all the datasets are essentially the same, the relationship is perfect and therefore the average loss value is zero. In such a case the fit attains its maximum value which equals the number of dimensions. In any other case, a fit value which is quite close to the number of new dimensions shows very good fitting of the correlation model obtained by the application of NLCCA to the original data. It is essentially a measure that denotes the capability of the new dimensions for describing the original data. For example, if the number of new dimensions is 50 as is the case for the correlation between the keywords and the MeSH terms and the average loss is 11.253 as described in the previous paragraph, then the fit value is 38.747 (see Section 4.2). This means that 77.5% (computed as a ratio of 38.747/50) of the original information is retained. If the correlation was perfect then in our example the fit value would have been 50.

The *eigenvalue* is a measure computed for each dimension and equals one minus the average loss for that dimension. Each eigenvalue indicates how much of the relationship is explained by the corresponding dimension. The sum of all the eigenvalues equals the fit value, so any eigenvalue can be expressed as the percentage of the total fit explained by the corresponding dimension. The eigenvalues are very useful in recognizing the most important dimensions that are able to describe adequately the data. High eigenvalues indicate an important dimension, whereas small ones signify that the dimension is not important and can be omitted. For example, in the application of NLCCA to the two matrices (keywords and MeSH terms) as in the previous examples, the first dimension has an eigenvalue of 0.927 and the fit is 38.747 (see Section 4.2), therefore $0.927/38.747 = 2.4\%$ of the actual fit is accounted for by the first dimension.

The *canonical correlation* is also computed for each new dimension by the corresponding eigenvalue using a straightforward formula: If E_i is the eigenvalue of

dimension $i = 1, ..., r$, the canonical correlation for the same dimension is:

$$\rho_i = \frac{[(K \times E_i) - 1]}{K - 1} \quad (3)$$

where K is the number of datasets in the analysis.

The projection obtained by NLCCA is essentially used for discrimination of the data points in homogeneous groups. The discriminating power of each one of the original variables, or else the importance of each variable for the data, is measured by a measure called *multiple fit*. In general, variables with higher values of the multiple fit discriminate better.

The results of NLCCA can be evaluated graphically by plotting in two dimensions the data points (in our case the documents) using as coordinates the scores of any pair of the new variables. These graphical representations are known as scatterplots [23,24], a simple but effective method to visualize combinations of numerical attributes. However, when the number of data points is very large, the plotting of all the data together causes problems in understanding relations between groups. In our case the groups are defined by the GO terms and one of the targets of the present study is to investigate whether the new variables computed by NLCCA can discriminate well, documents corresponding to different GO terms. So, in order to make the scatterplots more clear we used the so called *centroids*. The centroids are multidimensional points, computed for each GO term separately and they have as coordinates the means of all the new variables within the specific GO term/group. For example, suppose that 3 documents belonging to a specific GO term can be represented in a space of three dimensions by the coordinates given in Table 1. Then their centroid is a new three-dimensional point with coordinates $(0.5 + 0.3 + 0.4)/3 = 0.4$, $(1.2 + 0.4 + 0.2)/3 = 0.6$ and $(0.1 + 0.2 + 0.3)/3 = 0.2$. If the GO term has only these three documents assigned to it, we can represent the GO term in a three-axis Cartesian plot by the centroid point.

The plotting of centroids helps us to evaluate visually the results of NLCCA, since it depicts whether certain groupings are retained in the new projection space. In general, centroids that are placed far apart in a scatterplot, show distinguishable groups and therefore ability of the new dimensions to describe efficiently the groupings of documents.

Table 1
An example of the centroid for three documents belonging to a GO term

		Dimension 1	Dimension 2	Dimension 3
Document	1	0.5	1.2	0.1
	2	0.3	0.4	0.2
	3	0.4	0.2	0.3
	Centroid	0.4	0.6	0.2

4. Experimentation

The experimentation involves the application of NLCCA to $N = 9009$ biomedical documents [9] each represented by certain keywords, a number of MeSH terms and one GO term. The broad topics of the documents are "cell communication" and "cell growth" and they were retrieved using the e-utilities of Entrez.[4] Specifically, we modified the e-search and e-fetch utilities implemented by Oleg Khovayko[5] in order to submit queries to PubMed and retrieve the documents in XML format [2].

The queries submitted were relevant to $g = 12$ GO terms and contained keywords relevant to the GO terms (Table 3). The 12 GO terms are the same as the ones used in [9] and partly in [8] and [11], where they have been used to evaluate different classification methods. The whole procedure of first retrieving the relevant documents and then keeping only those assigned to a single GO term is described in detail in [9]. The purpose of using the same data as in our previous work [9] is because we have a good understanding of the data and is easier to test and evaluate new methodologies. Also, we have to point out that we chose to experiment with documents assigned with only one GO term, just for evaluation reasons. Indeed, if each document belongs to only one specific GO category then it is easier to evaluate the separation of the documents based on the new variables produced by NLCCA.

Regarding the size of the sample (9009 documents), it is a sample size consistent with those appearing in the related literature. Indeed, it is similar or larger than other corpora used in the field of biomedical text mining. For example, see [8, 20, 25, 26] and others.

In order to evaluate and compare the results to previous work we also used a test set of 8225 documents. The queries we used for retrieving these documents are the same as the ones for the 9009 documents (Table 3). The only difference is the date of publication. The 9009 ones were published up until 1999, whereas the 8225 between 2000 and 2004. The keywords of the test set are also the same as the ones in the training one. So, the test set is a 8225×1642 matrix.

The main goals of the experiments were:

1. To explore the correlation between the three different document representations.
2. To explore the trends, patterns and groupings of the data by graphically representing the documents in two-dimensional Euclidean spaces.
3. To represent the 9009 documents with a single matrix with a reduced number of variables, compared to the number of the original variables.

[4] "Entrez, The Life Sciences Search engine" is provided by National Library of Medicine (NLM) at National Center for Biotechnology Information (NCBI) of U.S. Government.

[5] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_example.pl.

4. To verify that NLCCA is capable to preserve the important information from the original data.

4.1. The three document representations

The first representation is a $(0,1)$-matrix with $N = 9009$ rows and $p = 1642$ columns corresponding to the total number of keywords extracted from the corpus of documents by the procedure described in Section 2.

The second representation is also a $(0,1)$-matrix with $N = 9009$ rows and $q = 50$ columns corresponding to the MeSH terms extracted from all the documents. The list of all the MeSH terms is given in Table 2. These are all major headings[6] [5] of the 9009 articles and each one of them is relevant to at least 100 articles, since the GO terms used in the experiments are also relevant to at least the same number of articles.

The third representation contains only one variable (i.e. a 9009×1 matrix) with $g = 12$ categories corresponding to the GO terms given in Table 3. It should be noted that three GO and MeSH terms are the same. These are: Cell Cycle, Signal transduction and Meiosis. Nevertheless, these terms are assigned to different documents. This is due to the fact that the assignment of the MeSH terms is done manually by the curators of the PubMed database, whereas the assignment of the GO terms is done using the query system of the PubMed database as described in Section 2.

Table 2
The MeSH terms extracted from the documents used in the experiment

MeSH terms		
Apoptosis	Cell cycle	Pharmacology
Cell cycle proteins	Cell transformation neoplastic	Physiopathology
Cell transformation viral	DNA damage	Toxicity
DNA methylation	DNA repair	Physiology
Drosophila proteins	*Escherichia coli* proteins	Radiation effects
Gene expression regulation	Gene expression regulation bacterial	Ultrastructure
Gene expression regulation neoplastic	Genes bacterial	
Genes fungal	Genes structural	
Genes tumor suppressor	Genes myc	
Genes p53	Genes ras	
Meiosis	Multigene family	
Mutation	Oncogenes	
Promoter regions genetics	Recombination genetic	
Saccharomyces cerevisiae proteins	Signal transduction	
Transcription genetic	Analogs 38 derivatives	
Analysis	Antagonists 38 inhibitors	
Biosynthesis	Chemistry	
Cytology	Drug effects	
Embryology	Enzymology	
Genetics	Growth 38 development	
Immunology	Metabolism	
Methods	Pathology	

4.2. Evaluation and interpretation of the results

The main issues examined in the results from the different experiments are:

- The strength of the correlation;
- The importance of each of the original variables for describing the documents;
- The descriptive power of the new variables;
- The graphical representations of the documents using the new variables.

We applied the NLCCA to different combinations of the representations. The results are described below.

4.2.1. NLCCA between the keyword and the MeSH term representations

The maximum number of new dimensions[7] calculated from these two matrices was $r = r_{\max} = 50$. The statistical measures obtained from the analysis are reported in Table 4 where for brevity the results for dimensions 11–47 are not shown.

The overall fit value (38.747) is quite high compared to the maximum number of dimensions ($r_{\max} = 50$) whereas the average loss (11.253) over all dimensions is quite low. This means that much of the original information about the documents is retained in the new variables. Moreover, the eigenvalues are high (0.927–0.704; Fig. 4) even for the last dimension, meaning that the new variables can describe a considerable portion of the variability in the data. We can also see from Table 4 that the correlation between the two document representations is significant at least for the first dimensions.

In conclusion, the 1642 binary variables for the keyword representation and the 50 binary variables for the MeSH terms can be replaced by a single dataset with 50 new numerical, real-valued variables which do not have any physical meaning, like the weights for keywords or the MeSH terms, but they are coordinates of points corresponding to documents in a new, lower dimensional Euclidean space. This new representation maintains a large percentage of the original correlation structure in the original spaces.

The box plots in Fig. 5 describe the distribution of the values of one of the new variables for each GO term. We can see that the documents corresponding to specific GO terms, like "cell death" and "oncogenesis", have distributions distinguishing them from the rest.

The importance of each of the original variables for describing the documents is assessed using the multiple fit measure in Tables 5 and 6. Table 5 gives basic descriptive statistics of the fit, whereas Table 6 shows the keywords and MeSH terms with the 10 highest multiple fit values. Overall, the MeSH terms have fit values above 0.5 while

[6] Describe the main topic of the documents.
[7] Each dimension corresponds to a new variable calculated by NLCCA.

Table 3
The GO terms used in the experiments and the queries used to retrieve them from PubMed

Number	GO code	GO term	PubMed query
1	GO:0006914	Autophagy	(Autophagy [TI] OR autophagocytosis [MAJR]) AND (Proteins[MH] OR Genes[MH]) AND 1940:1999[DP]
2	GO:0007049	Cell cycle	(Cell cycle[MAJR]) AND Genes[MH] AND 1996:1999[DP]
3	GO:0008283	Cell proliferation	(Cell proliferation[TI]) AND Genes[MH] AND 1940:1999[DP]
4	GO:0007267	Cell cell signalling	(Synaptic transmission[MAJR] OR synapses[MAJR] OR gap junctions[MAJR]) AND Genes[MH] AND1940:1999[DP]
5	GO:0006943	Chemimechanical coupling	(Contractile proteins[MAJR]) AND Genes[MH] AND 1993:1999[DP]
6	GO:0007126	Meiosis	(Meiosis[MAJR]) AND (Genes[MH] OR Proteins[MH]) AND 1986:1999[DP]
7	GO:0008152	Metabolism	(Metabolism[MAJR]) AND Genes[MH] AND 1989:1999[DP]
..			
8	GO:0007048	Oncogenesis	(Cell transformation, neoplastic[MAJR]) AND Genes[MH] AND 1994:1999[DP]
9	GO:0006950	Stress response	(Wounds[MAJR] OR DNA repair[MAJR] OR DNA damage[MAJR] OR Heat-Shock response[MAJR] OR stress [MAJR] OR starvation[TI] OR soxR[TI] OR (oxidationreduction[MAJR] NOT Electron-Transport[MAJR])) AND Genes[MH] AND 1996:1999[DP]
10	GO:0006810	Transport	(Biological transport[MAJR] OR transport[TI]) AND Genes[MH] AND 1985:1999[DP]
11	GO:0008219	Cell death	(Cell death[MAJR]) AND Genes[MH] AND 1997:1999[DP]
12	GO:0007165	Signal transduction	(Signal transduction[MAJR]) AND Genes[MH] AND 1995:1999[DP]

Table 4
Statistics from NLCCA applied to keywords and MeSH terms (summary of analysis)

		Average loss	Eigenvalue	Correlation	Fit
Dimension	1	0.073	0.927	0.854	
	2	0.106	0.894	0.788	
	3	0.121	0.879	0.758	
	4	0.128	0.872	0.744	
	5	0.137	0.863	0.726	
	6	0.139	0.861	0.722	
	7	0.149	0.851	0.702	
	8	0.157	0.843	0.686	
	9	0.159	0.841	0.682	
	10	0.169	0.831	0.662	
	...	...	...	...	
	48	0.292	0.708	0.416	
	49	0.295	0.705	0.410	
	50	0.296	0.704	0.408	
Sum		11.253			38.747

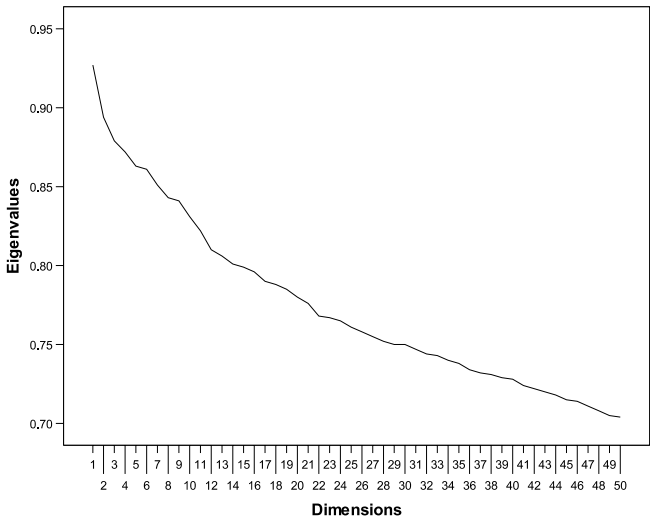

Fig. 4. The eigenvalues for each dimension of the NLCCA between the keywords and the MeSH terms.

the keywords below 0.5. This remark indicates that the MeSH terms can describe better the information contained in the biomedical documents.

Certain patterns and trends in the data using the new variables can be seen in Fig. 6 which is a scatterplot with axes corresponding to the first two new dimensions of the model. The numbered points represent the centroids of the article scores corresponding to the 12 GO terms. From this figure we can clearly see that only the first two dimensions obtained by the combined information from keywords and MeSH terms, are able to discriminate well the documents belonging to certain GO groups. For example, documents relevant to the 7th GO (''metabolism'') are clearly distinguished from all the others. Furthermore, documents that describe quite different biological phenomena, like ''transport'' (10th GO term) and ''cell death'' (11th GO term) are far apart, while others describing similar phenomena, like ''cell signaling'' (4th GO code) and ''signal transduction'' (12th GO code) are closer to each other. This verifies that although the original dimensions are reduced by NLCCA, the important information about the documents is retained. The use of more than the first two new dimensions can provide further insight to the information of the documents and distinguish them even better.

In order to test statistically whether each dimension separately depicts the grouping of the data, we used one-way parametric and non-parametric (Kruskal–Wallis) analysis of variance (ANOVA) tests [27] for all the new dimensions obtained by NLCCA. All tests gave significance $p < 0.0005$, i.e. there is always a significant difference between the groups defined by GO terms. The reason for performing those tests is to evaluate the discriminating power of the new variables. Indeed, we know that the original variables contain a large amount of information about the GO terms but since their space undergoes a huge distortion due to the dimensionality reduction by NLCCA, it is reasonable to

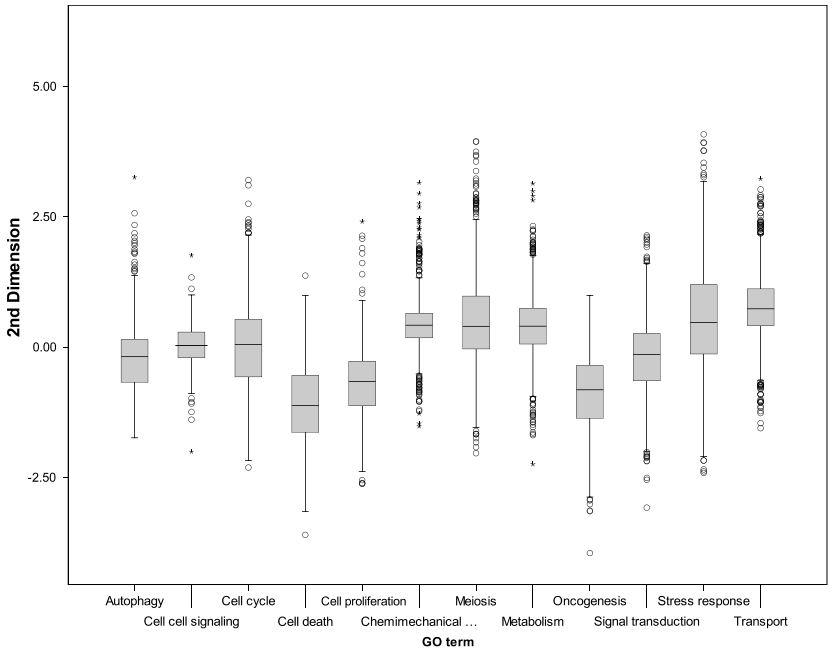

Fig. 5. A boxplot of the values of a new dimension for each GO term. The values resulted from NLCCA on the keywords and MeSH terms.

Table 5
Descriptive statistics of multiple fit for the original variables for all the NLCCA experiments

Descriptive statistics				
Documents representations	Minimum	Maximum	Mean	Standard deviation
Keywords and MeSH terms	0.009	1.497	0.0518	0.144
Keywords and GO terms	0	10.078	0.1	0.25
MeSH and GO terms	0.002	8.049	0.306	1.125
Keywords and MeSH and GO terms	0.002	8.049	0.306	1.125

Table 6
The 10 most descriptive MeSH terms and keywords based on NLCCA between the keywords and the MeSH terms

MeSH term	Multiple fit	Keyword	Multiple fit
Drug effects	1.497	p53	0.480
Pharmacology	1.366	Oncogene	0.425
Genetics	1.083	*Drosophila*	0.423
DNA methylation	0.961	Apoptosis	0.419
Meiosis	0.914	Repair	0.418
Metabolism	0.899	Methylation	0.365
Physiology	0.891	Damage	0.329
DNA repair	0.881	C-myc	0.291
Cell transformation neoplastic	0.879	Meiotic	0.284
Drosophila proteins	0.870	Recombination	0.227

wonder whether the new points (i.e. documents represented by new variables) keep their groupings or whether they are intermixed losing their initial information. So, the ANOVA tests show that the discriminating ability of the documents

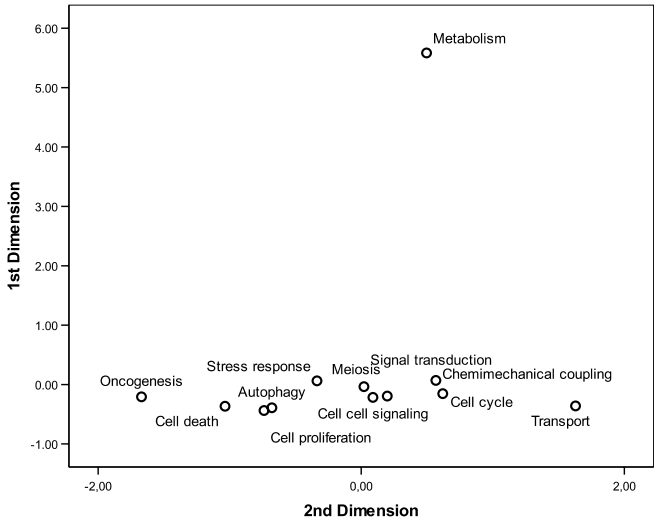

Fig. 6. Scatterplot of the GO centroids in the first two new dimensions for NLCCA between the keywords and the MeSH terms.

in the new low dimensional space is at least not completely lost. Later, in Section 4.2.1. we extend this evaluation using another one data set and more advanced multivariate statistical techniques.

Figs. 7–9 give the mean values of the document scores in the first three dimensions. The mean plots show that each one of the new variables can discriminate the documents describing at least one of the GO terms. For example in Fig. 7 we can see that the mean of document scores for the ''metabolism'' GO term is clearly different from the other means. Thus we verify that each new variable calculated by NLCCA can distinguish the articles and group them according to the GO terms they belong to.

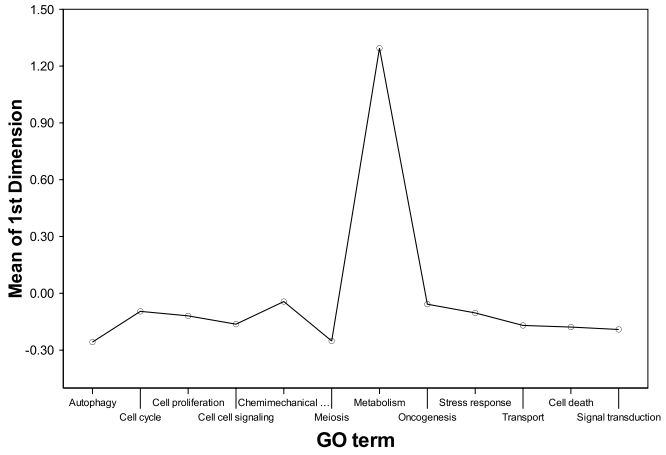

Fig. 7. Mean plot of document scores for 1st dimension.

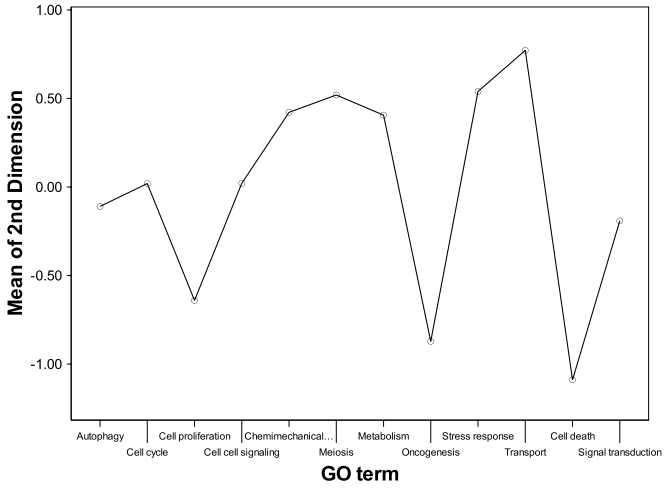

Fig. 8. Mean plot of document scores for the 2nd dimension.

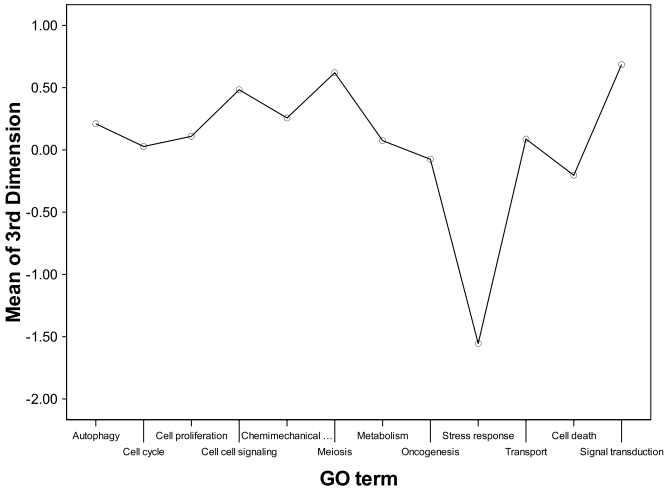

Fig. 9. Mean plot of document scores for 3rd dimension.

4.2.2. NLCCA between the keyword and the GO term representations

The number of dimensions computed was the maximum possible, $r = r_{\max} = 11$. Table 7 shows a summary of the measures obtained from the analysis. The small average loss (0.922) and the fit (10.078) indicate that there is very small amount of lost information from the original data. Furthermore, Fig. 10 depicts that the eigenvalues are high and the new variables are good descriptors of the variability in the data.

Table 7
Statistics from NLCCA applied to keywords and GO terms (Summary of analysis)

		Average loss	Eigenvalue	Correlation	Fit
Dimension	1	0.026	0.974	0.948	
	2	0.037	0.963	0.926	
	3	0.039	0.961	0.922	
	4	0.055	0.945	0.890	
	5	0.077	0.923	0.846	
	6	0.088	0.912	0.824	
	7	0.092	0.908	0.816	
	8	0.104	0.896	0.792	
	9	0.109	0.891	0.782	
	10	0.127	0.873	0.746	
	11	0.169	0.831	0.662	
Sum		0.922			10.078

Table 5 gives the basic descriptive statistics about the multiple fit of the original variables. Some variables have zero multiple fit, which means that they do not contain any important information for distinguishing the documents. These variables are keywords extracted from the documents. Table 8 describes the ten most descriptive (with highest multiple fit) original variables. The variable (GO indicator) that indicates the GO term that each document belongs to is the most important one with multiple fit equal to 10.078. The other nine variables are keywords describing certain biological processes or functions like the ones described by GO terms.

The scatterplot of Fig. 11 shows that the first two new dimensions contain sufficient information to distinguish documents relevant the to the GO term "Autophagy",

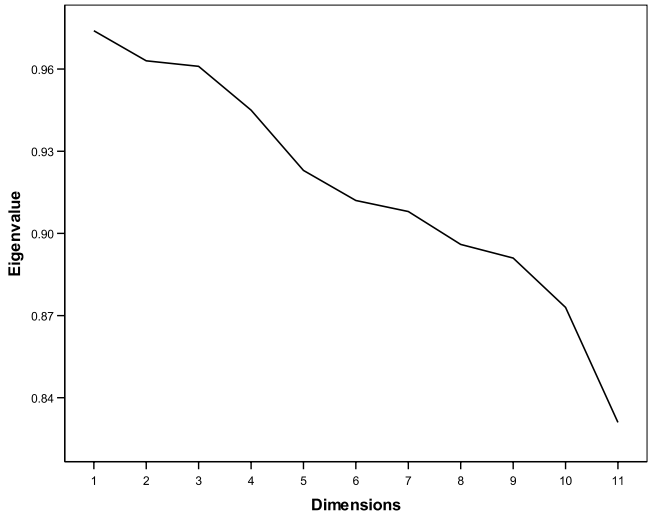

Fig. 10. The eigenvalues for each of the new dimensions of NLCCA between the keywords and the GO terms.

Table 8
The 10 most descriptive variables based on NLCCA between the keywords and the GO terms

Variable name	Fit value
GO_indicator	10.07800
Transportation	0.7730000
Proliferation	0.4680000
Apoptosis	0.4280000
Autophagy	0.3340000
Meiotic	0.2960000
Synaptic	0.2340000
Transformation	0.1470000
Repair	0.1430000
Meiosis	0.1040000

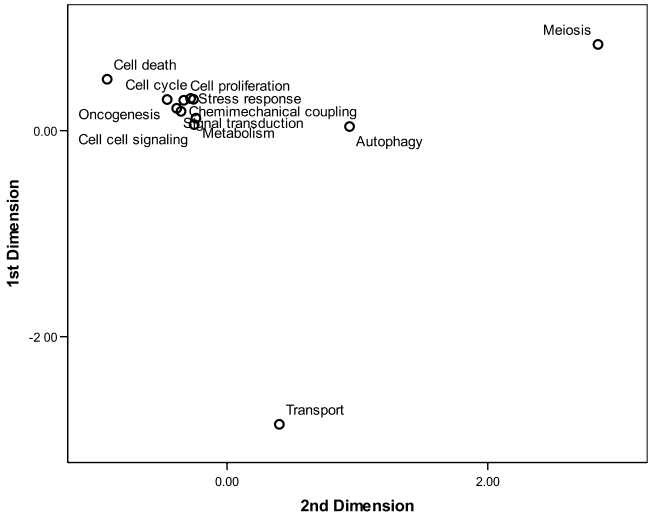

Fig. 11. Scatterplot of the GO centroids of the first two new dimensions for NLCCA between the keywords and the GO terms.

"Meiosis", "Transport" and "Cell death". However, documents describing irrelevant biological phenomena are far apart, like "Meiosis" and "Cell death". The ANOVA tests here gave the same results as in the previous datasets, i.e. each new variable shows significant difference between the GO codes.

Of course, here we have to emphasize that these results are more or less expected since the new variables contain the information of a grouping variable which is next used for visualization. However, what is important in this case is that we assess a high correlation between the GO terms and a data set of 1642 keywords. Furthermore, the low-dimensional space of only 11 variables is ideal for visualizing the documents in order to discover associations between GO terms, e.g. between "cell cycle" and "cell proliferation".

4.2.3. NLCCA between the MeSH and GO term representations

It is expected of course that there is some correlation between the MeSH and the GO terms. As can been seen from Table 9, the average loss values are higher here and the fit value of the model is not very high (8.049) since the number of dimensions is 11 ($r = r_{\max} = 11$). The eigenvalues (Fig. 12) and the correlations are comparable to the previous models for the first dimensions, but there are also dimensions with low values, (dimensions 8–11). Thus, although there is a certain degree of correlation between the two datasets, it seems that they describe the information contained in the biomedical articles differently.

Table 9
Statistics from NLCCA applied to the MeSH and GO terms (summary of analysis)

		Average loss	Eigenvalue	Correlation	Fit
Dimension	1	0.146	0.854	0.708	
	2	0.158	0.842	0.684	
	3	0.163	0.837	0.674	
	4	0.182	0.818	0.636	
	5	0.214	0.786	0.572	
	6	0.270	0.730	0.460	
	7	0.293	0.707	0.414	
	8	0.329	0.671	0.342	
	9	0.373	0.627	0.254	
	10	0.398	0.602	0.204	
	11	0.425	0.575	0.150	
Sum		2.952			8.049

Moreover, we can see from Tables 5 and 10 that the variable indicating the GO terms, as in the previous experiments, is the most descriptive one (multiple fit value = 8.049). The MeSH terms on the other hand are less important for describing the information of the documents (multiple fit values less than 1).

The scatterplot (Fig. 13) also shows that the articles can be distinguished using the first two new variable scores, especially those referring to "meiosis", "signal transduction" and "oncogenesis".

4.2.4. NLCCA between the keyword, MeSH and GO term representations

The information from the three matrices (keywords, MeSH and GO terms) was combined in order to produce

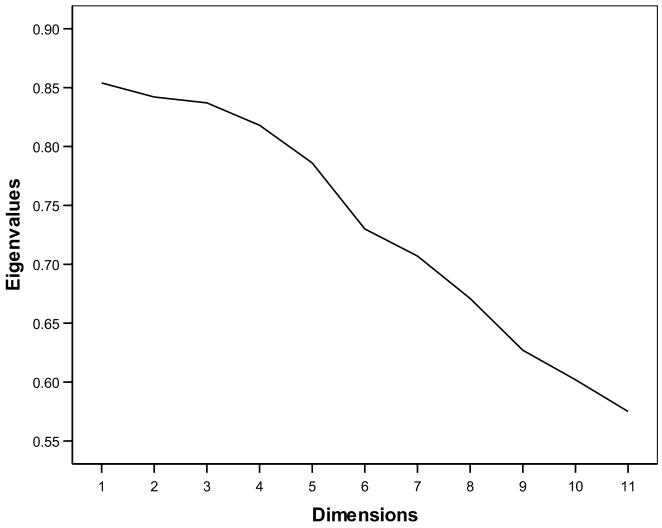

Fig. 12. The eigenvalues for the 11 dimensions of NLCCA between MeSH and GO terms.

Table 10
The first 10 most important variables based on NLCCA for MeSH and GO terms

Variable name	Fit value
GO_indicator	8.049
Signal transduction	0.75
Meiosis	0.723
Apoptosis	0.701
DNA methylation	0.62
Cell transformation neoplastic	0.507
Physiology	0.48
Cell cycle	0.469
Genetics	0.378
DNA repair	0.294

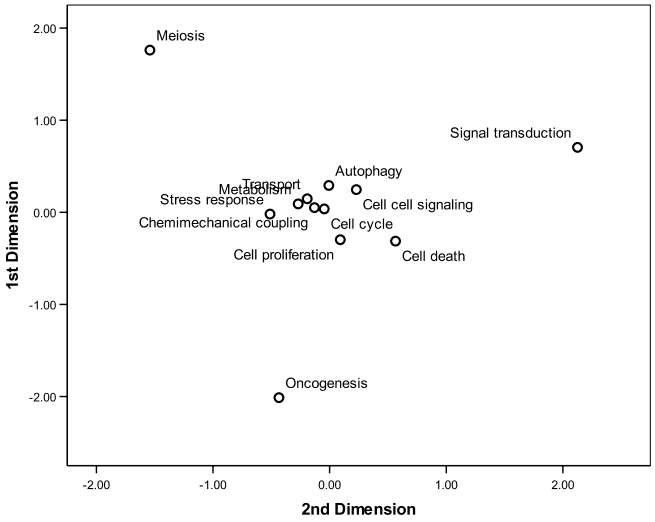

Fig. 13. The centroids of article scores of the first two dimensions for NLCCA between MeSH and GO terms.

a single matrix of new variables that could describe the documents. We calculated the first 20 variables ($r = 20$, $r_{max} = 1642$), because as can be seen from Table 11 and Fig. 14 after the twelfth dimension the eigenvalues and the correlations are too low and thus they describe only a small portion of the information of the documents and can be omitted. This is clearer in Fig. 14 where the steep of the line describing the eigenvalues changes at the twelfth dimension.

Tables 5 and 12 describe the 10 highest multiple fit values of the original variables (keywords, MeSH or GO terms) and indicate that the GO terms are the most important for describing the data. The MeSH terms and the keyword variables describe similar phenomena, but have different importance, i.e. ''Meiosis'' is a GO term and has a value of 0.601, whereas ''Meiotic'' is a keyword with value 0.285.

The scatterplot of Fig. 15 illustrates how well the first two dimensions can describe the information inside the documents and distinguish them according to the GO term they refer. Documents describing ''Meiosis'', ''Oncogenesis'' and the ''Cell death'' GO term are easily distinguished, whereas the differences between the rest of the GO codes, e.g. the ''cell cell signaling'' and ''signal transduction'' are also highlighted. In more detail, Fig. 16 depicts the difference in the values of the first two new dimensions for the documents belonging to the GO term ''Meiosis'' and ''Oncogenesis''. The ANOVA tests confirm the significant differences between the GO codes.

4.3. Evaluation of NLCCA with the use of classification techniques

In this section we evaluate further the ability of NLCCA to produce variables which can describe efficiently the information inside the documents. Especially, we are interested in evaluating through comparison with previous work, using the Linear Discriminant Analysis (LDA) and the Support Vectors Machines (SVM) classification methods. These classification experiments may be seen as an example of the practical usefulness of NLCCA to the biomedical domain. Our approach simulates the realistic situation where except from the documents for which we have a complete package of information including keywords, MeSH and GO terms, there are also new documents which have only the information from keywords (i.e. they have not yet been assigned MeSH or GO terms).

We performed two experiments based on the multiple fit results of NLCCA. The first one was based on the correlation between the keywords and the MeSH terms and the second one between the keywords, the MeSH and the GO terms. In the first phase of the experiments, the multiple fit computed by NLCCA was used to select the most descriptive keywords. Next, in the second phase, we used only these keywords to create an LDA and an SVM classification model. The two models were build using the 9009 older documents as the training data set. Then, the test set of the 8225 more recent documents was utilized to evaluate the classification results and compare them to previous work where all the keywords have been used [9]. In this way, the test set essentially simulates a corpus of plain documents waiting for classification into the 12 GO categories.

The multiple fit from the NLCCA between the keywords and the MeSH terms had a median value of 0.021. We selected all the keywords for the training and the test set of documents with a fit above the median. The number of keywords that were finally selected was 791.

After training the LDA model with the training 9009×791 matrix, we found the classification accuracy of the model on the test set (i.e. a 8225×791 matrix) to be 77.5%. The accuracy of the model when we used the training set as a test set itself (called fitting accuracy) was 87.1%. Finally, the accuracy using a hold-out sample was 72.46%. The hold-out sample is a random sample drawn from the training set, containing the 30% of the documents (3003). This subset is left out during the training phase of the model. The rest 70% of the documents (6306) were used for building the classification model. The corresponding results for the SVM classification model were 61.39% for

Table 11
Statistics from NLCCA applied to all the document representations (summary of analysis)

		Loss						
		Set 1	Set 2	Set 3	Mean	Eigenvalue	Correlation	Fit
Dimension	1	0.100	0.253	0.107	0.153	0.847	0.694	
	2	0.099	0.254	0.138	0.164	0.836	0.672	
	3	0.148	0.259	0.156	0.188	0.812	0.624	
	4	0.156	0.266	0.176	0.199	0.801	0.602	
	5	0.142	0.314	0.189	0.215	0.785	0.57	
	6	0.102	0.358	0.292	0.250	0.750	0.5	
	7	0.142	0.461	0.227	0.277	0.723	0.446	
	8	0.150	0.501	0.248	0.300	0.700	0.4	
	9	0.109	0.671	0.183	0.321	0.679	0.358	
	10	0.133	0.729	0.227	0.363	0.637	0.274	
	11	0.172	0.569	0.389	0.377	0.623	0.246	
	12	0.146	0.153	0.992	0.430	0.570	0.14	
	13	0.157	0.166	0.991	0.438	0.562	0.124	
	14	0.154	0.224	0.955	0.444	0.556	0.112	
	15	0.164	0.234	0.953	0.450	0.550	0.1	
	16	0.169	0.228	0.955	0.451	0.549	0.098	
	17	0.180	0.201	0.980	0.454	0.546	0.092	
	18	0.187	0.209	0.986	0.460	0.540	0.08	
	19	0.197	0.199	0.998	0.465	0.535	0.07	
	20	0.199	0.231	0.969	0.467	0.533	0.066	
Sum		3.003	60.482	11.111	60.865			13.135

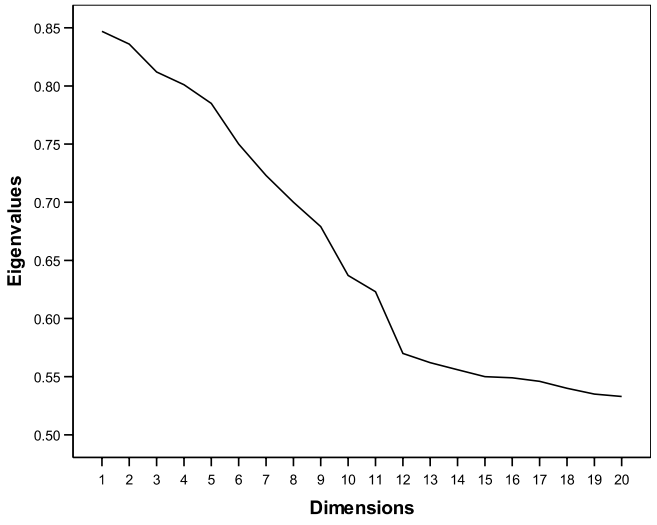

Fig. 14. The eigenvalues of the 20 new dimensions of NLCCA using the three document representation.

the test set, 99.84% for the fitting accuracy and 64.74% for the hold-out sample.

Similarly, the multiple fit computed by NLCCA for the correlation between the keywords, the MeSH and the GO terms had a median of 0.003 and the keywords selected above 0.003 were 790. The classification results of the LDA were for this experiment 78.44% for the test set, 87.55% for the fitting accuracy and 73.36% for the hold-out sample. The corresponding SVM results were 69.33% for the test set, 99.82% for the fitting accuracy and 66.7% for the hold-out sample.

It is interesting to see the classification results when using all the 1642 keywords in the experiments. For LDA we have 77.31% classification accuracy for the test set, 88% fitting accuracy and 80.81% classification accuracy for the hold-out sample. For SVM, the classification accuracy for the test set is 71.99%, the fitting accuracy is 97.49% and the classification accuracy for the hold-out sample is 73.8%.

Table 12
The first 10 most important variables based on NLCCA between the three document representations

GO term variable	mF value	MeSH terms	mF value	Keywords	mF value
GO_indicator	8,88,900	DNA Methylation	0.819	Transport	0.640
		Immunology	0.721	Apoptosis	0.449
		Radiation effects	0.717	P53	0.353
		Signal transduction	0.629	Prol	0.318
		Meiosis	0.601	Meiotic	0.285
		Apoptosis	0.583	Autophagy	0.284
		Genes ras	0.568	Repair	0.279
		DNA Repair	0.563	*Drosophila*	0.279
		Genes p53	0.521	Methylation	0.263
		Embryology	0.482	Synaptic	0.202

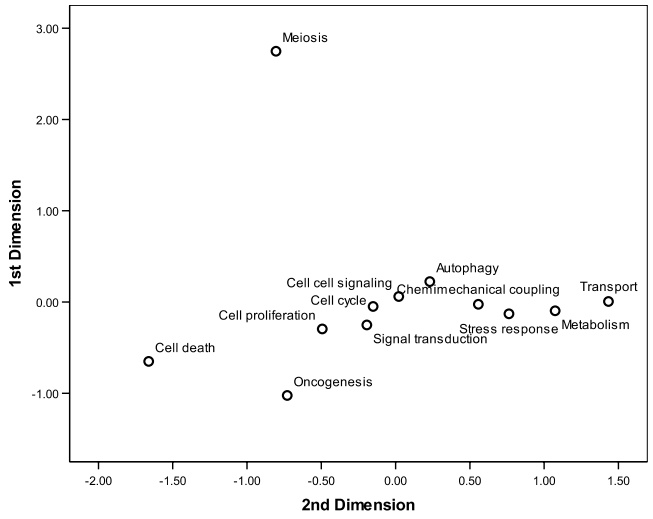

Fig. 15. The scatterplot of the centroids of the article scores based on NLCCA between the three document representations.

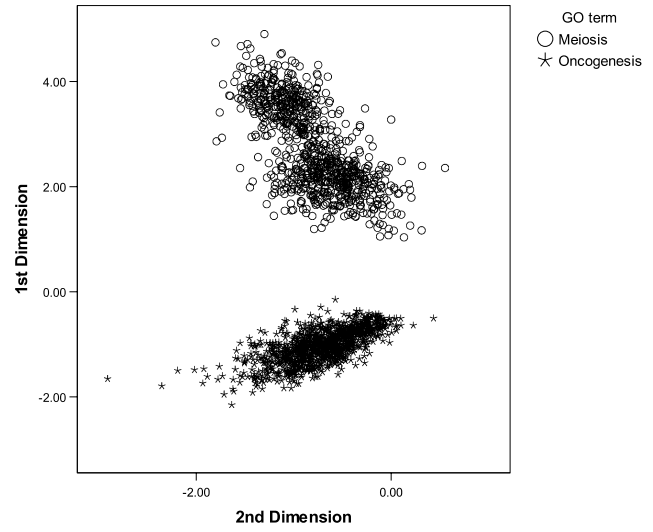

Fig. 16. A scatterplot of the documents relevant to GO term "Meiosis" and "Oncogenesis". It is clear that the documents are perfectly discriminated.

All the results are summarized in Table 13. It is obvious that the measures of NLCCA provide valuable information about the importance of the keywords and can help us to keep and work with only the most descriptive keywords. In our case, the variable selection based on the fit values reduced the number of keywords from 1642 to around 790, i.e. less than the half. The selected, through NLCCA, keywords provided enough information to build efficient LDA and SVM classification models.

Finally, we have to emphasize here the twofold advantage of using NLCCA. In the previous sections we saw that NLCCA can map the original data by appropriate transformations in a new low-dimensional space, useful for data visualization and exploratory analysis. In this section we evidenced that NLCCA can also serve as a means for variable selection, useful in classification.

5. Conclusions and future work

The results from our experiments indicate that the new variables of the NLCCA model provide useful information about the data and help us clearly to distinguish the documents based on the GO term they describe. NLCCA can also significantly reduce the number of variables contained in the three datasets, for example in our experiments the 1642 keywords and the 50 MeSH terms are reduced to only 50 new variables. Furthermore, NLCCA facilitates the visualization in two dimensions of our multidimensional datasets in order to discover patterns and specific trends in graphical plots of the data. Another interesting conclusion indicated by the multiple fit values of NLCCA is that the keywords do not describe the documents as well as the MeSH or GO terms.

It should be noted that the NLCCA can be used as the first phase of a clustering or classification procedure such as SVM, naïve Bayes, discriminant analysis, etc [15]. The new variables can be further utilized in building classification models in order to categorize data. The interesting point here is that the new variables are uncorrelated and since many classification or clustering algorithms assume that the variables of the datasets are uncorrelated we can clearly see the advantage. Note also that in the case of keywords we have original variables highly correlated to another one, since each word depends on other words of the same text. Therefore, NLCCA can remove this correlation which might be a problem for building a model.

The fact that NLCAA can combine information from more than two different datasets makes it a very good

Table 13
The classifications results using the NLCCA information (correlation of keywords–MeSH and of keywords–MeSH–GO) for selecting keywords for training the models

Datasets used for classification						
Measures of classification accuracy	Keywords selected from NLCCA on keywords and MeSH terms		Keywords selected from NLCCA on keywords, MeSH and GO terms		All keywords	
	LDA (%)	SVM (%)	LDA (%)	SVM (%)	LDA (%)	SVM (%)
Fitting	87.1	99.84	87.55	99.82	88	97.49
Hold-out	72.46	64.74	73.36	66.7	80.81	73.8
Test set	77.5	61.39	78.44	69.33	77.31	71.99

The last two columns have the classification results when all the keywords are used.

method for biology, where there is usually the need to combine information from different data, like for example microarrays, biomedical articles, sequence and structure data. For example, NLCCA could be used in whole genome analysis to describe genes with one dataset. It could utilize information from the text, the MeSH and GO terms, the phylogenetic profiles, etc. of the genes. The new dataset could then be used into a clustering method in order to extract useful biological knowledge, such as pathways, interactions, etc. Another application of NLCCA could be in the field of health care and medical informatics. An important aspect in the "post genomic" era is the correlation between the genotype and the phenotype [28]. NLCCA could be used for combining the clinical data of the patients with their genotype in order to group them into different categories and discover specific trends and patterns for a disease.

As part of our future work we would like to apply NLCCA to describe genes and proteins using information from different sources, like text, expression profiles, etc. in order to discover novel biological knowledge. Also, we would like to apply NLCCA in the field of text clustering and more specifically in the field of biomedical article categorization. We would like to introduce a new methodology that could combine different sources of information in order to improve the efficiency and the usefulness of existing clustering methods, like K-means clustering.

Acknowledgment

We thank the anonymous reviewers whose comments and suggestions led us to improve significantly the paper.

References

[1] Rubin DL, Thorn CF, Klein TE, Altman RB. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. J Am Med Inform Assoc 2005;12(2):121–9.
[2] PubMed. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institute of Health (NIH). http://www.pubmed.com.
[3] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nat Genet 2000;25(1):25–9. URL: http://dx.doi.org/10.1038/75556.
[4] Consortium GO. Creating the gene ontology resource: design and implementation. Genome Res 2001;11(8):1425–33.
[5] Bean CA, Green R, editors. Relationships in the organization of knowledge. NY: Kluwer Academic Publishers; 2001.
[6] Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings versus text word with PubMed. Laryngoscope 2006;116(2):336–40.
[7] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 2006;22(6):658–64.
[8] Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res 2002;12 (1):203–14. URL: http://dx.doi.org/10.1101/gr.199701.
[9] Theodosiou T, Angelis L, Vakali A, Thomopoulos GN. Gene functional annotation by statistical analysis of biomedical articles. Int J Med Inform 2007;67(8):601–13.
[10] Lee M, Wang W, Yu H. Exploring supervised and unsupervised methods to detect topics in Biomedical text. BMC Bioinform 2006;7(1):140.
[11] Izumitani T, Taira H, Kazawa H, Maeda E. Assigning gene ontology categories (Go) to yeast genes using text-based supervised learning methods. In: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004); 2004, p. 503–4.
[12] Yasunori Yamamoto, Toshihisa Takagi. Biomedical knowledge navigation by literature clustering. J Biomed Inform 2007;40(2):114–30.
[13] Salton G. Automatic text analysis. Science 1970;168:335–43.
[14] Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003;10(6):821–55. URL: http://dx.doi.org/10.1089/106652703322756104.
[15] Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge, MA: The MIT Press; 1999.
[16] Michailidis G, De Leeuw J. The Gifi system for descriptive multivariate analysis. Statistical Science 1998;13:307–36.
[17] van der Burg E, De Leeuw J, Verdegaal R. Non-linear canonical correlation analysis. Leiden: Int Rep 1984:RR12–=0?>RR84. Department of data theory.
[18] Jurafsky D, Martin JH. Speech and natural language processing. Engelwood Cliffs, NJ: Prentice Hall; 2000. ISBN: 0-13-095069-6.
[19] Paice C. Another stemmer. SIGIR Forum 1990;24(3):56–61.
[20] Iliopoulos I, Enright AJ, Ouzounis CA. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. Pac Symp Biocomput 2001:384–95.
[21] Zobel J, Moffat A. Exploring the similarity space. SIGIR Forum 1998;32(1):18–34.
[22] Meulman JJ, Heiser WJ, SPSS Categories User's Manual v. 14.0, SPSS Inc.
[23] Lattin J, Carroll D, Green R. Analyzing multivariate data. Curt Hinrichs; 2003.
[24] Hair JF, Anderson RE, Tatham RL, Black WC. Mutlivariate data analysis. 5th ed. Englewood Cliffs, NJ: Prentice Hall; 1998. ISBN: 0-13-930587-4.
[25] Shah Parantu K, Carolina Perez-Iratxeta, Peer Bork, Andrade Miguel A. Information extraction from full text scientific articles: where are the keywords? BMC Bioinform 2003;4:20.
[26] Hong Pan, Li Zuo, Vidhu Choudhary, Zhuo Zhang, Houi Leow Shoi, Teen Chong Fui, et al. Association miner: a system for exploring transcription factor associations through text-mining. Nucleic Acids Res 2004;32(Web server issue):W230–4.
[27] Weiss NA. Introductory statistics. 6th ed. Reading, MA: Addison-Wesley; 2002. ISBN 0-201-71059-5.
[28] Martin-Sanchez F, Iakovidis I, Nørager S, Maojo V, De Groen P, Van der Lei J, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform 2004;37(1):30–42.